New Solutions for Navigating an Ocean of Data

S. M. Mesick
NOAA National Coastal Data Development Center
Stennis Space Center, MS 39529 USA

P. S. Hill, S. T. Gottfried
General Dynamics Information Technology /
NOAA National Coastal Data Development Center
Stennis Space Center, MS 39529 USA

***Abstract*- With more than 90 percent of the world's ocean still unmapped and unexplored, the need for ocean exploration has never been more critical. NOAA's Office of Ocean Exploration and Research (OER) provides NOAA and the Nation with a unique capability to discover and investigate new ocean areas and phenomena and to conduct the basic research required to capitalize on discoveries.**

In 2002 NOAA's Office of Ocean Exploration (now part of OER) initiated a joint effort with NOAA's Data Centers[1] and other partners to form an Integrated Product Team (IPT) for Data Management for the Ocean Exploration Program. The IPT researched and documented a strategic approach, and has since implemented an "End-to-End" (E2E) Information Management System to ensure that the scientific data and value-added data products produced as a result of NOAA's exploration program are appropriately managed.

The cornerstone of the E2E System is the Cruise Information Management System, an open source, custom software suite designed to aggregate information collected from OER's expedition planning and operational processes into standard documentation formats (e.g., metadata records). In step with the new sensor suites and technical capacity available aboard the NOAA Ship *Okeanos Explorer* (EX), CIMS capabilities are currently being extended to address new data management challenges resulting from the new technologies aboard ship. An important objective is to automate the creation of standard metadata records for a myriad of shipboard sensors with a minimum of human intervention. Other key technical investigations include the transformation of shipboard sensor data collections to open standards formats to enable near-real time data access and automated archival, as well as investigation into the use of a shore-side Data Assembly Center to provide a common framework for data transformation and distribution.

The team's approach to systems development, emphasizing collaboration, flexibility, adaptation, and transparency, remains on course to meet future expedition information management needs. The IPT ensures that the information resulting from OER's global, interdisciplinary expeditions is broadly accessible to decision makers, scientists, educators, and the public, and is preserved for perpetuity. OER will not only serve NOAA's present needs, it will undoubtedly bring to light what will become of NOAA's and the Nation's future missions and priorities.

[1] This work is jointly funded by OER and NOAA's Data Centers.

I. INTRODUCTION

The past several decades have brought significant changes to physical, chemical, and biological ocean environments. With more than 90 percent of the world's ocean still unmapped and unexplored, the need for exploration has never been more critical. Built from the merger of two unique NOAA programs— NOAA's Undersea Research Program (NURP) and the Office of Ocean Exploration (OE),— NOAA's newly formed Office of Ocean Exploration and Research[2] (OER) is poised to build on a rich legacy of undersea exploration, discovery, and research. OER will provide NOAA and the Nation with a unique capability to discover and investigate unexplored ocean areas and phenomena, will conduct the basic research required to capitalize on discoveries, and will seamlessly disseminate scientific data and value-added data products (information) to a multitude of users.

The initial creation of OE in 2001 marked NOAA's response to the Report of the President's Panel on Ocean Exploration [1]. OE's objectives sought to focus the best undersea assets and ocean scientists' minds on conducting reconnaissance expeditions to investigate unknown and poorly known ocean areas. From the outset, an essential element of the OE program has been information management and dissemination.

Following recommendations from the President's Panel, OE began an extramural collaboration with NOAA's Data Centers and other partners to "Establish a broad-based task force to design and implement an integrated, workable, and comprehensive data management information processing system for information on unique and significant features" [1, p. 44].

This collaboration quickly formalized into NOAA's Integrated Product Team (IPT) for Data Management for the Ocean Exploration Program. Initially the IPT focused

[2] NOAA OER is mandated by the Omnibus Land Management Act of 2009, Public Law 111-11 Title XII: Oceans, Subtitle A: Ocean Exploration, Parts I and II, Section 12001 – 12107.

0-933957-38-1 ©2009 MTS

on documenting requirements and assessing both partner and community capabilities and tools available to meet the requirements. A data management strategy for the OE Program was developed and documented. [2]. The IPT then identified gaps between requirements, management objectives and available capacity and prioritized these for action. Two systems were prototyped, one for video management and once for geospatial data visualization, analysis and access. Beginning with the 2006 OE field season, an information management system was loosely knitted together and tested under the rigors of real-world data collection [3]. Over time, prototypical systems became operational and were integrated with other system components, forming an End to End (E2E) Information Management System. The original requirements, focusing solely on meeting information management needs for OE's annual scientific field season, are now managed operationally.

Exploration information management requirements have continued to evolve since the original assessment was completed. The NOAA Ship *Okeanos Explorer*[3] (EX), "America's Ship for Ocean Exploration," fields a variety of sensors, data collection, and transmission systems that advance a new paradigm for exploration. These exciting scientific advances also represent a new paradigm for exploration information management, increasing both the amount and type of information to manage, as well as the opportunities for automation, standardization and dissemination. The OER merger also presents potential opportunities for centralization of selected information management functions. Changing standards within the information management community (i.e., new Federal metadata standards) and new technologies that are widely accepted by end users (such as Google Earth™) also impact information management requirements.

The IPT Executive Committee continues to assess changing requirements and to systematically prioritize tasks to meet evolving organizational information management requirements. Each year's Annual Operating Plan (AOP) reflects IPT guidance, and assists IPT Working Groups (WG) in annually refocusing activities and resources to meet requirements.

II. E2E PROCESS: BACKGROUND AND HIGHLIGHTS

The primary goals of the E2E system are to ensure broad accessibility to, and preservation of, the sound scientific data and value-added data products (information) resulting from ocean explorations. From the outset, IPT members realized that the common thread between the immediacy of collecting and processing information during an expedition, and the long-term goals of providing public access to information and preserving it for perpetuity, is standard documentation (i.e., metadata) [4].

[3] Learn more about the NOAA Ship *Okeanos Explorer* at this web site: http://explore.noaa.gov/technology/okex.

The IPT assessment revealed a series of procedural steps, were routinely executed by OE staff to move an approved, funded proposal to a state of post-expedition completion. While these procedures were not part of a seamless process, and did not generate standards-based metadata, the similarity in information content between existing procedures and standard metadata formats was notable. It seemed evident that using existing steps to produce an additional product (metadata records) was both a straightforward scenario to implement and to accomplish with minimal organizational impact.

OE's expedition management procedures were documented and diagramed, then streamlined to propose a workflow that would result in an E2E system; that is, a system that manages information from proposal through archival and makes information readily accessible. Software tools used by IPT collaborators were evaluated, such as the Management Information System (MIS) used within NURP for proposal management, and the Expedition Information System (EIS) used within OE to document at-sea scientific data collection. A common data model was created, which borrowed from MIS, EIS and other resource materials. This complex data model was "mapped" to the standard metadata model, documenting a path for OE to produce the requisite documentation.

The IPT crafted a set of guiding principles to move this plan forward toward an operational system, as follows:

- streamline and automate expedition planning and operational procedures, produce standard documentation;
- adopt and adapt existing data management tools;
- use open-source standards for maximum efficiency and transparency; and
- sustain the collaboration between OER's exploration program and its many partners, drawing on those pools of expertise and resources.

Ultimately none of the existing tools met the overarching requirements, so a new software system was designed. The resulting system, called the Cruise Information Management System (CIMS), forms the cornerstone of the information management system. Shown in Fig. 1, CIMS enables data discovery, access, and preservation.

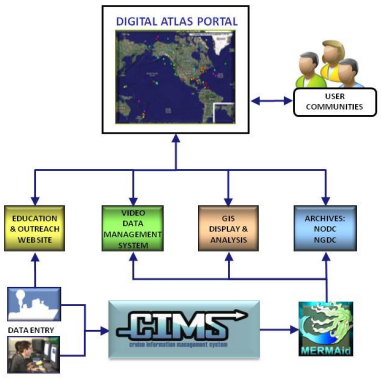

Figure 1. CIMS is the cornerstone of the E2E system

CIMS is a web-based data entry system that was built to open standards using license-free software. CIMS modules streamline and automate the procedural workflow to produce standard metadata records in compliance with government mandates and community standards. Operationally, completed metadata records are bundled with scientific data and value-added data products for archival and to ensure broad discovery and access to information. Ongoing CIMS development activities focus on increasing automated information throughput from shipboard sensors to distributed destinations.

Specific tools have also been developed to facilitate user community access to NOAA's exploration information. The Digital Atlas Portal[4] is an easy-to-navigate Google™ map application that displays expedition locations on a global map. Through the Digital Atlas, the user community may directly download scientific data and value-added data products from distributed data repositories, including NOAA archives, NOAA Library catalogs, and geospatial databases. Geographic Information System (GIS) tools are available to visually integrate and analyze geospatial information.

The Video Data Management System (VDMS) provides an innovative solution to the challenge of managing NOAA's collection of expedition videos, images, and supporting documents (i.e., cruise plans, situation reports, final cruise reports, and similar). The VDMS holdings may be accessed directly through the Digital Atlas, where users can stream video clips and exploration highlight videos, as well as view images and documents [5].

While software automation and cool web tools can facilitate reaching information management goals, software alone cannot address the complexity of getting an expedition off the desktop and out on the ocean. A large network of people make an expedition "Go!", and the IPT working groups play a key role in assuring that information management is included in the earliest stages of expedition planning. OER data managers are CIMS users; they also train others to use CIMS. They provide expedition support in the office and on the sea, following up with expedition scientists post-cruise to ensure that information is preserved and is accessible in a timely manner, while at the same time ensuring that sensitive resources are protected. They affect the necessary cultural and organizational changes required to make systems like CIMS, VDMS and the Digital Atlas useful, and E2E information management a reality.

III. THE IMPORTANCE OF METADATA

As stated, standard metadata records form the common thread that connects user communities to scientific data and value-added data products, and makes this information usable now and in the long term. A metadata record is a file that captures basic scientific data characteristics, such as where and when scientific data was collected and with what instrumentation. When scientific data undergoes processing (i.e., for quality control, for data analysis or product creation), metadata records are updated with process steps to ensure continued accuracy. Developing metadata to meet Federal standards, as well as community guidelines, enables discovery and interoperability.

The Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata[5] (CSDGM) is mandated for use in documenting Federal geospatial data collections. This standard has been implemented in the CIMS as part of the operational E2E system. References within standard metadata records link information about the data to individual datasets for direct access. Metadata records that are published to discovery portals, such as the Geospatial One Stop,[6] enable the public to discover and access a wide array of geospatial information.

Other significant standards are also utilized in the E2E system. The Library of Congress MAchine-Readable Cataloging[7] (MARC) format makes up the foundation for documenting most bibliographic information found in library catalogs. Librarians at the NOAA Central Library worked with IPT partners to extend the MARC standard to accommodate documentation of video and still images [5]. When metadata records are published in the extended MARC formats, information may be discovered through the Library's online catalog NOAALINC.[8]

Other important standards guidance is provided by the IOOS DMAC Committee [6], particularly for data broadcast and access protocols. The E2E system complies with DMAC guidance as much as possible, particularly in designing and implementing software solutions to document, process, and archive observed shipboard sensor data collections.

Creating detailed and accurate metadata can be a time-consuming process, which places demands upon limited data processing resources. When planning, the IPT sought to reduce the overhead associated with metadata record creation by automating processes to the fullest extent possible.

As discussed, the information about an expedition generated by OER's expedition management workflow strongly aligns with many of the FGDC standard metadata elements. IPT partners "mapped" the content between OER's administrative

[4] The Digital Atlas is an index to OER's exploration data collection. Direct links to data archives, library holdings and geospatial data and tools are provided. http://www.ncddc.noaa.gov/website/google_maps/OE/mapsOE.htm.

[5] The U.S. Office of Management and Budget and the U.S. Congress set policy for Federal agencies. OMB Circular A-16a defines the FGDC responsibility to prepare and maintain a strategic plan for the development and implementation of the National Spatial Data Infrastructure. The FGDC develops geospatial data standards for implementing the NSDI in accordance with OMB Circular A-119. The FGDC website is a resource for governance, policy and standards information. http://www.fgdc.gov/.

[6] geodata.gov is a geospatial data portal, also known as the Geospatial One-Stop (GOS) that is maintained by the FGDC to serve as a public gateway for improving access to geospatial information and data under the Geospatial One-Stop E-Government initiative.

[7] The Network Development and MARC Standards Office at the Library of Congress and the Standards and the Support Office at the Library and Archives Canada maintain the MARC 21 formats. Input for development is provided by MARC 21 users from around the world, including libraries, library networks and utilities, and library system vendors. See http://www.loc.gov/marc/overview.html for more information about the development and maintenance of the formats.

[8] Information about the NOAA Central Library and access to Library collections through NOAALINC. http://www.lib.noaa.gov/.

documents and the FGDC standard metadata format. CIMS software was designed to gather this information in a step-by-step manner consistent with the ongoing workflow. Additional mapping between the FGDC and MARC standards [7] has extended this approach such that maximum utility is gained from any data entry task.

CIMS's modular design mimics this procedural workflow, with the goal of reducing duplicative data entry tasks, and potential errors as well. The catch phrase is to "take the pain out of metadata creation"; this is accomplished by aggregating information entered into the system as it follows the workflow.

IV. CIMS OVERVIEW

A. *CIMS Modular design*

The CIMS software design initially focused on the development of modular components closely corresponding to discrete elements of OER's long-standing procedures for managing OE's annual scientific field season expeditions. In this model, scientists have an annual opportunity to submit proposals to OE; cruise plans are generated to develop awarded proposals into sea-going expeditions; once expeditions are underway, data collection activities are centrally managed to produce standard documentation files; post cruise, information is fully accessible for enhancement and modification. The CIMS design has been enhanced to meet additional requirements to manage scientific data and value-added data products in near-real time. The CIMS Broker component addresses this requirement. Fig. 2 provides an overview of the modular design, color-coded to show the status of each module's development. In this figure, orange striped boxes indicate that the modules are not fully developed, but alternative methods enable at least partial requirements to be met; Green striped boxes indicate software modules in operational beta mode; solid green boxes indicate fully operational software.

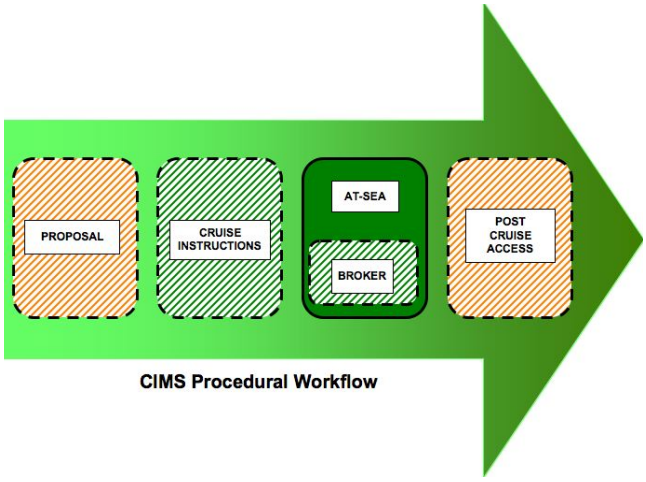

Figure 2. CIMS is a modular software system

B. *Proposal Management*

OER issues annual Announcements of Opportunity (AO) that result in funding for selected ocean exploration activities. Proposals for funding may be submitted to OER via Grants Online,[9] and must include an OER cover sheet that provides an overview of the full proposal. The initial CIMS software design planned for the development of a secure, online data entry tool to gather proposal information, initialize CIMS, and pass information along for field season planning (i.e., internal review and cruise planning) in a "tactical decisions aid" format.

The operational Grants Online website, which was developed in the same time frame as the initial CIMS design, clearly supersedes the need to develop the CIMS Proposal Module as originally planned. Currently the information needed to initialize CIMS is gleaned from the OER Cover Sheet and manually loaded into CIMS.

The IPT is gathering additional, incidental requirements related to proposal management. While the proposal is still an important source of metadata content, it is no longer ubiquitous to the CIMS process (i.e. proposals are not associated with EX operations). Development plans may be impacted by the need to (1) develop investment metrics and tracking; (2) receive information and/or report out to other information management systems; and (3) integrate and streamline OER administrative activities.

C. *Cruise Instructions*

The recently developed CIMS Cruise Instructions (CI) module is a secure, web-based data entry system designed to enable groups of expedition principals to collaboratively build detailed operational plans for OER expeditions. Using this tool, participants enter expedition-specific information about the personnel, targeted exploration locations, planned science activities, vessel, instrumentation, and other factors that make up a particular expedition. Information entered into CIMS as part of this process is also collected into a database populated through use, which allows the system to "learn" or develop a set of reference information that will be available for subsequent / future plan development. This information will be an invaluable asset in preparing the CIMS At-Sea module for use. Information entered into the CI module builds a set of operational instructions (i.e. a Cruise Plan) that is an important record of the planning process and that will be archived with the other cruise materials. The CIMS CI module is currently in beta testing, and use is limited to the EX operations team.

D. *At-sea data collection*

The CIMS At-sea module was prioritized for early development and implementation by the IPT. As such, it has the longest operational history and is the most widely used element of CIMS. It is a platform-independent, license-free portable software suite that can operate on a shipboard

[9] Grants Online is the Federal solution for full life-cycle grants management processing. For more information visit https://grantsonline.rdc.noaa.gov/flows/home/Login/LoginController.jpf.

computer (from various locations via the ship's network) or on a standalone system.

The CIMS At-sea data entry tool allows direct initialization of the system prior to use. In this scenario, OER data managers pre-load the system from a written, expedition specific, data management plan. Once initialized, the CIMS At-sea data entry screens present users with the planned schedule of shipboard activities scheduled throughout the expedition. Data managers can record information about shipboard activities in real time. CIMS At-sea produces FGDC CSDGM standard compliant files in XML format.

The current release, CIMS At-sea V1.0, was built primarily for use during the annual field season, when expeditions take place on ships of opportunity or are sometimes shore based. It uses the original, IPT defined data model and technology stack.

The CIMS Broker is a new software component that has been developed to facilitate direct CIMS integration with shipboard systems and sensors for metadata creation and data transformation. The next release of the At-sea module will integrate the Broker component.

CIMS V2.0 is scheduled for release in FY10 and will include significant upgrades. Most notably, CIMS 2.0 will be built around a newly integrated data model, which will harmonize all the operational CIMS modules into a unified structure.

E. Expedition Portal

Another component of the original design is the Expedition Portal, designed to provide access to the expedition's metadata records for post-cruise management. The idea was that authorized expedition participants could access records to complete tasks and to perhaps generate "child" records with a minimum of duplication. Much of the requirement for post-cruise metadata record management is now met by the integration of CIMS-produced records into the Metadata Enterprise Resource Management Aid[10] (MERMAid), a freely available, versatile metadata management tool developed at NOAA's National Coastal Data Development Center (NCDDC). Authorized staff may have access to expedition metadata records, and may utilize all the metadata record management capacity of the MERMAid system.

The expedition portal was also planned to provide OER expedition metrics, such as number of dive operations performed in a year, number of mammals sighted, or estimations of square kilometers of ocean floor mapped. The need for metrics is ongoing, and a system to easily generate this information is still planned. Requirements will be reviewed prior to design to ensure that all needs are captured and that systems already in place are fully utilized. In particular, the system must seek to unify proposal and performance accountability for accurate tracking.

F. CIMS Architecture

The initial guidance to use open-source standards for maximum efficiency and transparency has had a positive impact on the CIMS development to date. Two specific elements have contributed to the overall success of the software system.

The first element is the selection of the Python programming language. This open-source language offers terse yet elegant software solutions, and generally requires less coding to accomplish complex tasks than other comparable software languages. This reduces both development time and code maintenance. Python also provides a plethora of libraries and utilities that have been extremely useful in developing CIMS. Python does not have a steep learning curve, thus lending itself to ready collaboration.

The second key element was the decision to utilize open standards formats for information files. This decision has been primarily beneficial in the use of XML for metadata output and for passing information between modules. Recently, the use of open standards formatting has been expanded to include the Hierarchical Data Format[11] (HDF5) binary data file format. Using these and other open standards:

- makes it easier to collaborate with other developers when designing and implementing the exchange of data between CIMS and external systems;
- extends the development network to include tools developed by others for data discovery and access;
- allows data to be used interoperability with other NOAA and non-NOAA data for visualization, analysis and decision support.

V. Creating Metadata with CIMS

A unique instance of CIMS At-sea is created for each expedition. Each instance is initialized by information gleaned from the full proposals and from cruise instructions as they are finalized. The OER data manager uses the CIMS At-sea instance on the expedition, replacing paper data logs, spreadsheets and other tools with the CIMS digital interface.

The CIMS At-sea User Interface (Fig. 3) is an easily modifiable daily planner style calendar that is initialized from the cruise instructions detailed itinerary. As the expedition progresses, the OER data manager can document expedition operations, activities, and events. Calendar entries that produce, record, or otherwise generate information for archival can be updated with the specific documentation about that event.

[10] The Metadata Enterprise Resource management Aid (MERMAid) is a license-free, web-based tool used to develop, validate, manage, and publish metadata records via secure internet access. For more information visit http://www.ncddc.noaa.gov/metadataresource/metadata-tools/view.

[11] **Hierarchical Data Format**, commonly abbreviated **HDF**, **HDF4**, or **HDF5** is the name of a set of file formats and libraries designed to store and organize large amounts of numerical data. It is currently supported by the nonprofit HDF Group, whose mission is to ensure continued development of HDF5 technologies, and the continued accessibility of data currently stored in HDF. Visit http://www.hdfgroup.org/ for more information.

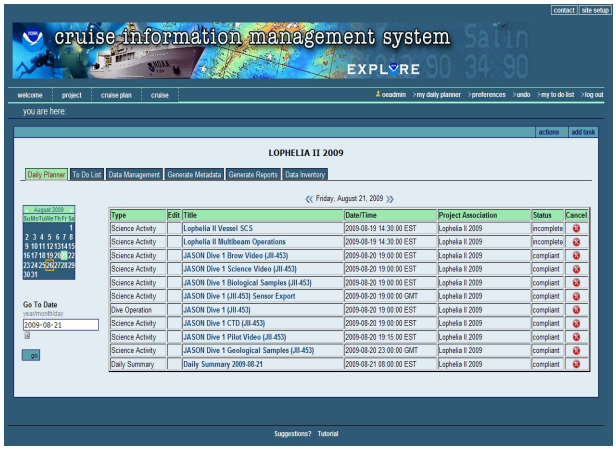

Figure 3. The CIMS At-sea module user interface

The calendar can be modified to reflect real-time changes to planned activities caused by weather events, equipment failures or opportunities to pursue more detailed investigation of a given phenomena

Post cruise, the CIMS-produced CSDGM XML formatted files are ingested into MERMAid, validated for FGDC standards compliance, published to GOS and bundled with scientific information for archival at NOAA Data Centers.

MERMAid enables expedition principals to access and manage metadata records post cruise. FGDC metadata records generated from CIMS may be modified within MERMAid. Records may be:

- updated to document additional data processing steps;
- copied and modified to create new records for value-added data products;
- exported as XML files; and
- exported in MARC compliant XML for Library management.

This versatility maximizes the resources invested in creating standard documentation. This approach enables OER data managers using CIMS to create metadata records in multiple formats, meeting more than one documentation requirement with consolidated data entry actions.

VI. A NEW PARADIGM FOR INFORMATION MANAGEMENT

The IPT continues to evolve data management planning in step with the new sensor suites and technical capacity available aboard the *Okeanos Explorer*. Working in collaboration with NOAA's Office of Marine and Aviation Operations (OMAO), CIMS is being adapted to read files generated by the shipboard Scientific Computing System [7] (SCS).

NOAA vessels typically transmit subsets of shipboard sensor data periodically from the ship to a variety of shore-side data customers, such as the NOAA ship tracker database (hourly), the National Oceanographic Data Center (NODC) Shipboard Sensor Data Base (various intervals dependent on ship capabilities), and others. The complete data collection from a given shipboard operational period may also be copied to portable media for distribution to the NOAA Chief Scientist for any given activity.

While each of these data processes may meet unique user requirements, there appears to be no single method for ensuring standardized documentation, timely access to information in open standard or community accepted formats, or routine preservation of NOAA shipboard sensor data collections to all appropriate repositories.

Further, there is an ineffective redundancy in the processes necessary to generate and transmit each separate data package. Software systems require maintenance and attention from shipboard crewmembers to operate; transmission to various shore-side destinations requires repeated use of sometimes limited bandwidth from the ship's internet connection. Further, various formats transmitted from ship to shore are subject to interruption or incomplete trans-missions based upon the level of data compression, as well as the quality of the shipboard connection. Other resources on the receiving end of the data stream are required to ensure that each data package makes it to the intended recipient both on time and complete. Overall, these combined methods are inefficient, and produce unnecessary overhead to accomplish routine data processing and transmission tasks.

The IPT's Software Development Working Group (SDWG) was challenged to develop a new data management paradigm to manage the unique data collection and transmission capabilities aboard the EX. Opportunities to streamline, automate and standardize shipboard data management activities have been identified. To meet these new requirements, the CIMS software has been enhanced with an additional component called the CIMS Broker ('Broker').

The Broker performs three primary functions specific to shipboard sensors and systems: (1) automating the creation of FGDC standard metadata files; (2) transforming scientific data files from shipboard sensors and systems to open standard formats; and (3) transmitting these files to a shoreside component for management.

The SCS, operational aboard NOAA ships including the EX, monitors and records shipboard system activities and events, and also records scientific data for a defined suite of standard shipboard sensors and systems. Other shipboard systems, such as the EX's hull mounted EM302 multibeam system, record scientific data files on shipboard computer systems external to the SCS (for purposes of discussion these will be referred to as 'internal' and 'external' sensor systems, respectively).

To create metadata for internal shipboard sensors the Broker reads information directly from SCS generated, fixed format files. The software methods and information workflow are similar to those used to create the CIMS/CSDGM files with CIMS At-sea V1.0, with the primary difference being the methods of data entry. To create metadata for external shipboard sensors, the Broker reads both SCS and external sensor data files, combining key information from multiple sources to complete the CSDGM template. Once CIMS is initialized, the Broker processes for creating CIMS/CSDGM files are fully automated and do not rely on data managers to

log shipboard activities into the system. Shipboard operations logged by SCS are also captured by the Broker. The operational beta version of the Broker continues to rely on MERMAid to validate and publish the CIMS/CSDGM files. In CIMS V2.0, this dependency will be removed.

The CIMS Broker transforms scientific data files collected from internal shipboard sensors and recorded by the SCS, into the open standard HDF5 format. The HDF5 format was selected because of the ability to store and organize large amounts of numerical data, and because it is an efficient format for data transport. The SCS can also monitor and record scientific data collected with additional sensor systems such as over-the-side CTD rosettes and ROV mounted video systems. The Broker framework is extensible, and the SDWG is developing algorithms to transform each additional sensor format to HDF5 on a case-by-case basis.

One HDF5 file is created daily for all internal sensor data collected for any given day. The Broker automatically prepares and transmits HDF5 files from the ship to NCDDC's Ecosystem Data Assembly Center (EDAC). Files are transmitted from ship to shore daily and contain all the scientific data collected by the SCS for that day (Fig. 4). In addition to the HDF5 data file, documentation and value-added data products may also be part of the transmission. Software services in the EDAC verify file integrity and perform several subsequent operations.

HDF5 files are collected within the EDAC for additional post-cruise processing. When the expedition is completed, HDF5 files are further processed to create individual NetCDF[12] files for each data variable for the entire cruise. The entire data collection for the cruise, as well as the individual variable files will be transmitted to NODC for access from the NODC OPeNDAP[13] server and for archival.

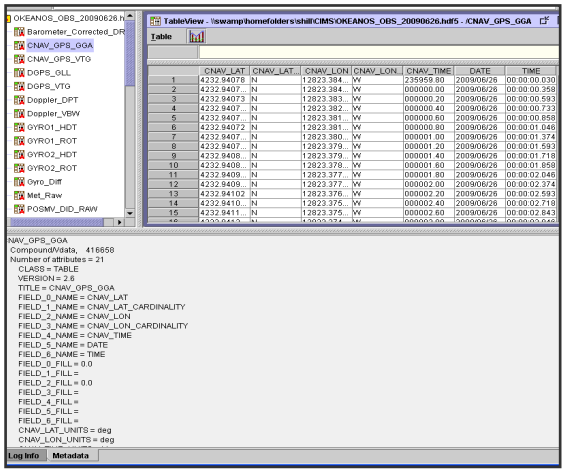

Figure 4. The CIMS Broker transmits files from ship to shore daily

Once this method is fully operational, HDF5 files aggregated at the EDAC will leverage the full operational capacity of the NCDDC Regional Ecosystem Data Management (REDM) services architecture. REDM services that may prove useful for OER data management include automated bundling (data with metadata) and distribution to NOAA archives at prescribed intervals, user subscription to routine data feeds in user-defined formats via syndication services, public access to data and data products via web visualization tools and the EDAC THREDDS servers.

An important note is that the EX multibeam files are not transformed into HDF5. OMAO survey technicians perform initial multibeam processing, some analysis and product development on board ship. While some mapping products may be part of the CIMS Broker daily transmission to NCDDC, the data files are not part of the daily transmission. Post-cruise, multibeam files are transferred to the National Geophysical Data Center (NGDC) for archival, and, by special arrangement between OER and the University of New Hampshire, to the UNH Center for Coastal and Ocean Mapping[14] for more sophisticated processing and analysis. Because of the size of these files, a rotational hard drive system is used to transfer the data between locations. As the EX telepresence system becomes fully integrated into shipboard operations, elements of described data transport methods will evolve.

Using Google map the automated technologies described herein, IPT members at NCDDC have prototyped a map that provides near-real time public access to EX cruise tracks, bathymetric data and mapping products. Information to populate the Okeanos Explorer Digital Atlas[15] (Fig. 5) comes from the daily file transmissions.

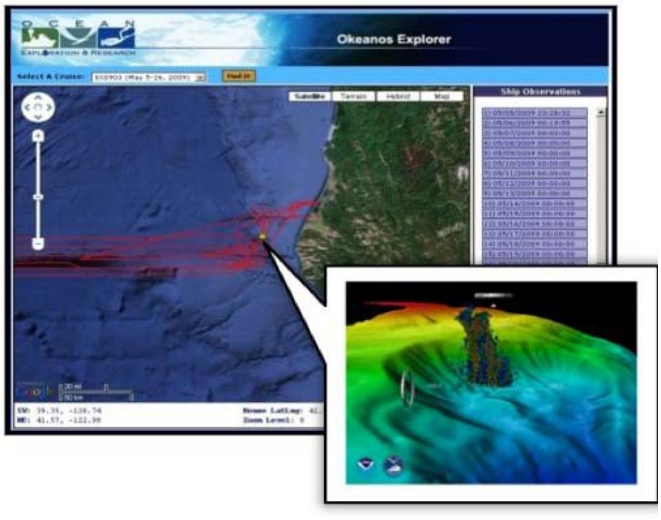

Figure 5. The Okeanos Explorer Digital Atlas provides public access to cruise information and value-added data products in near-real time

[12] http://www.unidata.ucar.edu/software/netcdf/

[13] OPeNDAP is a framework that simplifies all aspects of scientific data networking. For more information visit opendap.org. The NODC OPeNDAP server is available from www.nodc.noaa.gov/opendap

[14] http://ccom.unh.edu/

[15] http://www.ncddc.noaa.gov/website/google_maps/OkeanosExplorer/mapsOkeanos.htm

The CIMS Broker is currently installed on the EX in a beta test capacity. As the EX moves forward with sea trials and testing of various shipboard sensors, SDWG members travel aboard ship and work collaboratively in real time with shoreside reach-back teams to test the data transformations and transmissions, and to further automate the enhanced E2E workflow. With planning, these initiatives hold the promise to reduce the burden on the crew for uploading, emailing, or otherwise transmitting the same information to numerous recipients. Routine automation of the transmittal of shipboard sensor data to NOAA's data centers will ensure that NOAA's investment in data collection is protected, and that data is preserved and accessible in standard interoperable formats in near-real time.

VII. Future plans

Each fiscal year the IPT develops an AOP to define the work plan for that period. All information management tasking, including CIMS development, is annually reprioritized based upon the greatest need and the overall resource allocation. The overall priority is to make CIMS fully operational on board the EX. However the best plans cannot account for every lesson learned and applied during sea trials. Enhancements and completion of other modules have taken a secondary role to the demands of real-time development and processing. The IPT Working Groups strive to adapt and adopt in response to evolving priorities and changing schedules.

For example, OER Proposal Cover Sheets are formatted in XML and can be used (with manual data entry methods) to directly initialize CIMS with pertinent information. While this approach may not meet all the CIMS Proposal Management software development requirements, this method enables the team to utilize existing information and technology to meet the specific need to initialize CIMS for metadata creation.

Similarly, CIMS CI requirements developed by potential system users represented a detailed list of desired capabilities for secure user roles and responsibilities, user interface forms and report formatting options and a reliance on a controlled vocabulary not yet available. The beta CIMS CI is a bare bones versions; the next release is planned to meet user requirements adjusted with 'lessons learned' from working with the beta version. Participants (e.g. expedition scientists) will engage more directly in information management by sharing the data entry workload based on system roles. As the CI is used, the common vocabulary and database will continue to grow, allowing users to:

- Select metadata keywords from predefined metadata keyword vocabularies, allowing scientists to enhance data discovery by common search engines and linking scientific data more closely to publications and other post-cruise materials;
- Indicate the types of data that will be collected and select where and when the data will be archived, greatly enhancing data preservation efforts;

The CIMS Broker will continue development as additional sensors, systems and technologies come online aboard the EX, The SDWG will continue to work with the NCDDC and the EDAC architecture teams to enhance automation, data throughput and public access to information. The Broker will be fully integrated into the CIMS At-sea module and will be self-contained in the capability to validate FGDC CSDGM standard metadata records independent of MERMAid.

The CIMS V2.0 will be released in Fiscal Year 10, and will represent a significant improvement that will integrate CIMS operations across all modules. The new release will be built around an enhanced data model that has been evolving for some time as real world use and additional functions (particularly related to EX sensor suites and technologies) have been tested and implemented. The new release will codify these changes.

The At-sea module will have a new user interface, with updates also based on real world experience; the Broker will be fully integrated into the At-sea module. Implementation of standard vocabularies will contribute toward simplification of manual data entry and will increase standardization. CIMS V2.0 will provide internal metadata validation capabilities, reducing reliance on connectivity. The CIMS V2.0 software suite will also undergo a technology refreshment to meet evolving NOAA information technology (IT) security standards.

The IPT will continue to improve methods of information delivery to a broad, multi-user community. Where appropriate, geospatial data will be accessible in KML format, and may become available via subscription in user-defined formats. This approach ensures that NOAA's exploration data are interoperable with other geospatial data for decision support.

VIII. Summary

In practice, the E2E system continues to adapt to meet evolving data management challenges:

- changing data formats and media (i.e., from standard to high definition video);
- new information technologies (such as remote science and cloud computing);
- changes to national data management standards and policies (such as FGDC adoption of the North American Profile);
- enhanced IT security profiles.

These challenges also present new opportunities, which the IPT will continue to prioritize and address while continuing to manage NOAA's exploration information. Each additional sensor added to the EX's suite provides an opportunity for enhancing data throughput. The OER merger also potentially adds the need to manage additional types of information, such as that associated with research.

The team's approach to E2E information management, which emphasizes flexibility, adaptation, and transparency,

remains on course to meet future ocean exploration and information management needs. The IPT ensures that the sound scientific data and value-added data products (information) that results from OER's global, interdisciplinary expeditions are broadly accessible to decision makers, scientists, educators, and the public, and are preserved for perpetuity. OER will not only serve NOAA's present needs, but also will undoubtedly bring to light what will become NOAA's and the Nation's future missions and priorities.

ACKNOWLEDGMENT

The authors wish to recognize the contributions of the IPT Executive Committee and Working Group members for their many contributions and unflagging commitment to the key tenets of data management, which have made this work possible, meaningful and enjoyable.

The authors also wish to thank NOAA managers within OER, NODC and NGDC for providing the resources that make the work possible.

REFERENCES

[1] "Discovering earth's final frontier: a U.S. strategy for ocean exploration: the report of the President's Panel on Ocean Exploration." Boulder, Colorado. University Corporation for Atmospheric Research, 2000. Available online, last accessed on August 21, 2009. http://explore.noaa.gov/media/http/pubs/pres_panel_rpt.pdf.

[2] *Data Management Strategy for the Ocean Exploration Program*. 2002. Silver Spring, MD, National Oceanic and Atmospheric Administration, Office of Ocean Exploration. (Contract number: 50-SPNA-9-00009; Task order: 56-SPNA- 9-90020; April 30, 2002)

[3] S, Mesick, S. Gottfried, and E. Roby, 2005. *Ocean Exploration Data Management: A Success Story*. In: *Oceans 2005 MTS/IEEE, Washington D.C.. Track: Information Processing, Modeling, and Simulation, Session: IPMS-7 Ocean Exploration Data Management.* Columbia, Md.: Marine Technology Society; Piscataway, N.J.: IEEE. CD-ROM. Available online, last accessed on August 21, 2009. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=1640065&isnumber=34367.

[4] *Ocean Exploration Program Metadata Guidelines.* 2002 Silver Spring, MD, National Oceanic and Atmospheric Administration, Office of Ocean Exploration. (Contract number: 50-SPNA-9-00009; Task order: 56-SPNA- 9-90020; November 25, 2002)

[5] A. Fiolek, D. Anderson, D.W. Collins, M.L. Cumberpatch, and J. Beattie, 2005. *NOAA Video Data Management System (VDMS): No Video, No Oscar®!* In: *Oceans 2005 MTS/IEEE, Washington D.C.. Track: Information Processing, Modeling, and Simulation, Session: IPMS-7 Ocean Exploration Data Management.* Columbia, Md.: Marine Technology Society; Piscataway, N.J.: IEEE. CD-ROM. Available online, last accessed on August 21, 2009. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=34367&arnumber=1640067&count=485&index=341

[6] *Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems Interoperable Data Discovery, Access, and Archive*, unpublished. Draft plan (dated March 2005) available online, last accessed August 21, 2009. http://dmac.ocean.us/dmacPlan.html.

[7] A. Fiolek, S. Ladnier, S. Gottfried, and S. Mesick, 2009. *Mapping, Cross-walking, Converting and Exchanging Oceanographic Metadata Information in Video Data Management System.* In: *Oceans 2009 MTS/IEEE, Biloxi, MS: 8.0 MARINE LAW, POLICY, MANAGEMENT, AND EDUCATION: 8.8 Marine education and outreach Principal Tech Area*, in press.

[8] T.Stepka, D. Shields, S. Chung, J. Katebini. *Scientific Computer System (SCS) Version 4.1.7 User's Guide*, unpublished. Silver Spring MD. NOAA OMAO Systems Group.